



**Barcelona
Supercomputing
Center**

Centro Nacional de Supercomputación



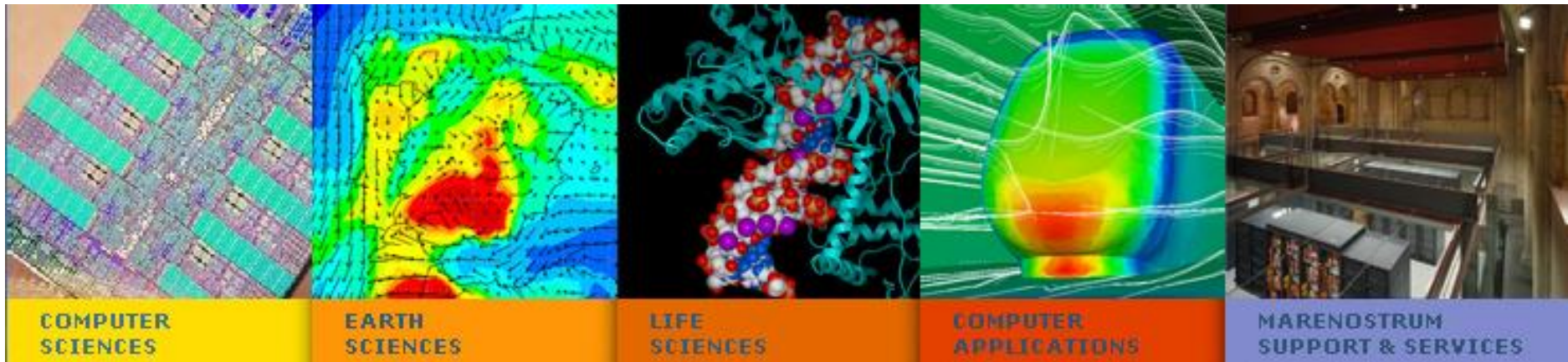
EXCELENCIA
SEVERO
OCHOA

Big Data en el estudio del clima y la calidad del aire

XVII Encuentro nacional de estudiantes de
matemáticas

Kim Serradell



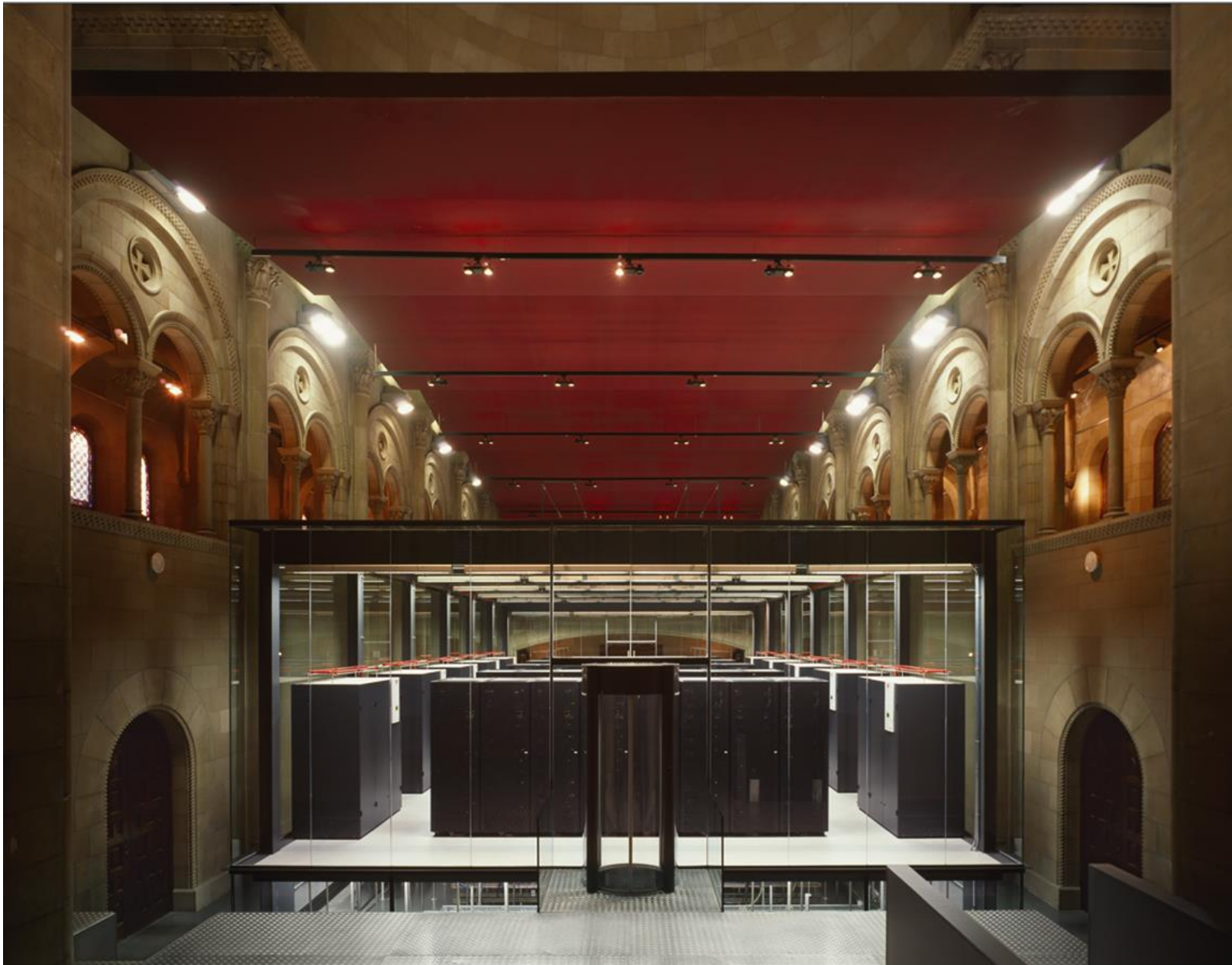


Mare Nostrum 3



**Barcelona
Supercomputing
Center**
Centro Nacional de Supercomputación

EXCELENCIA
SEVERO
OCHOA



Qué

Prógnostico de lo que nos rodea

Porqué

Nuestros puntos fuertes...

... investigación...

... operaciones...

... servicios...

... alta resolución...

Cómo

Desarrollar una capacidad para modelar procesos de calidad del aire de escala urbana a global y los impactos sobre el clima, la salud y los ecosistemas

Implementar un sistema de predicción del clima “seasonal to decadal”

Desarrollar servicios orientados al usuario que favorezcan tanto la transferencia como la adaptación de tecnología

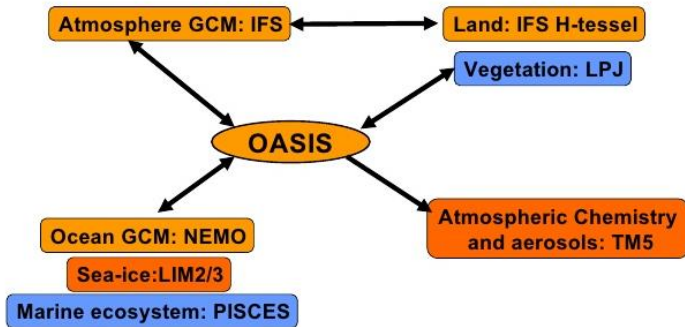
Utilizar HPC y tecnologías Big Data para la eficiencia y la facilidad de uso de los modelos de ciencias de la tierra

Earth system
services

Climate
prediction

Atmospheric
composition

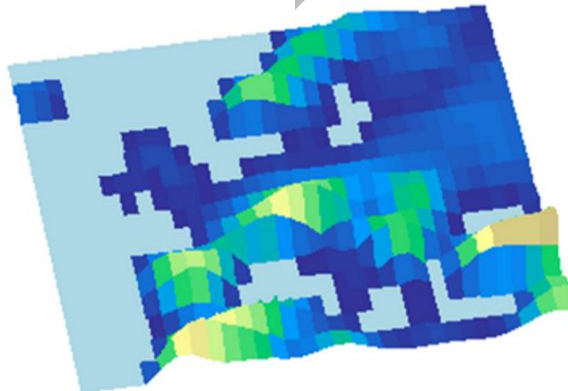
Computational
Earth sciences



- EC-Earth climate model
 - IFS para la atmósfera y NEMO para el océano
 - Acoplado online con OASIS y XIOS para el servidor de escritura
- Financiado por PRIMAVERA (H2020) y propuestas MINECO
- Recursos computacionales en MN3 (PRACE), ECMWF y ARGONNE (INCITE)

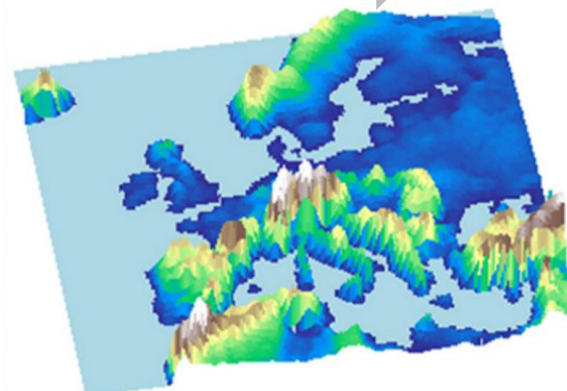
Resolución habitual
en los experimentos
actuales

1° – 100km



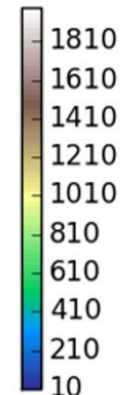
Resolución en
desarrollo por varios
centros a nivel mundial

0.25° – 40km



BSC y otros tres
institutos a nivel
mundial

0.12° – 10km

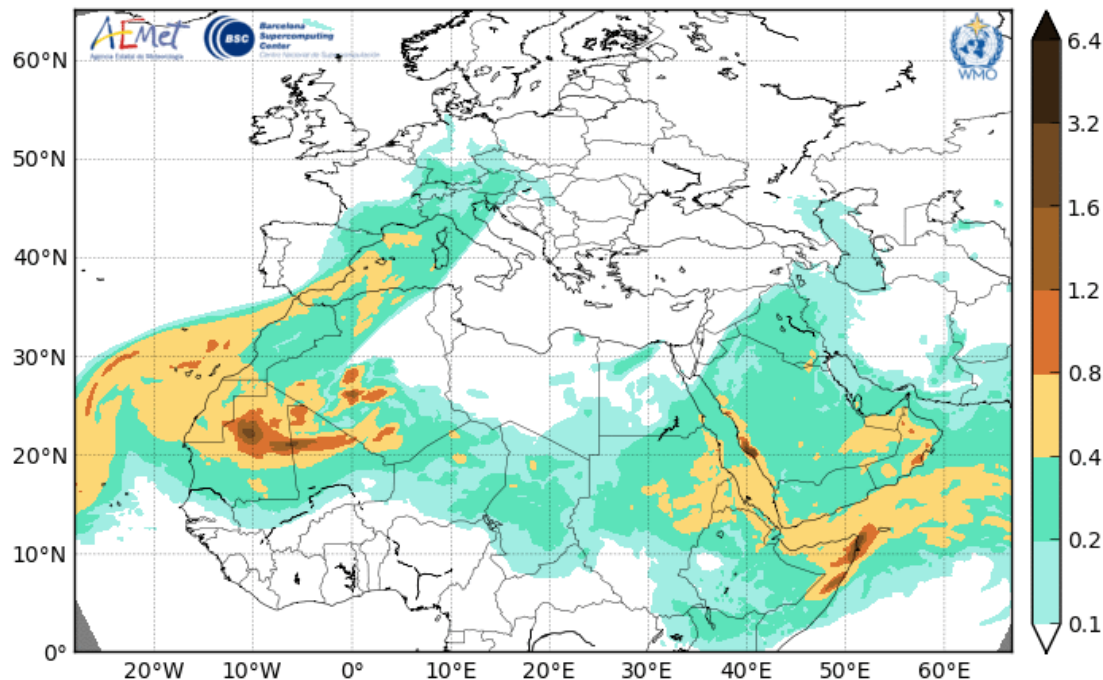


- *Nonhydrostatic Multiscale Meteorological Model on the B-grid*
 - Desarrollado en colaboración con el National Centers of Environmental Prediction (USA)
 - En el BSC se han desarrollado los módulos de aerosoles y química en fase gas

Barcelona Dust Forecast Center - <http://dust.aemet.es/>

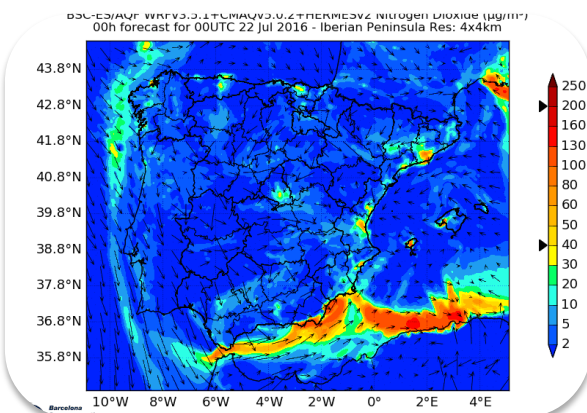
NMMB/BSC-Dust Res:0.1°x0.1° Dust AOD

Run: 12h 21 JUL 2016 Valid: 12h 21 JUL 2016 (H+00)

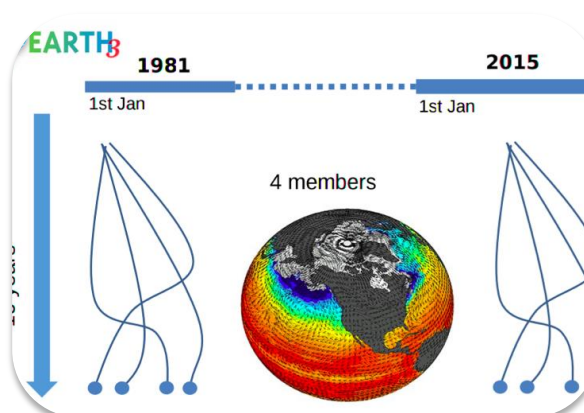


EC-Earth	Resolución horizontal (atmósfera/océano)	Tamaño de las salidas para un año de simulación (en NetCDF, ficheros de restart no incluidos)
Resolución estándar	T255/ORCA1 60km/100km	26 Gb
Alta resolución	T511/ORCA025 40km/25km	120 Gb
Muy alta resolución	T1279/ORCA012 25km/12km	1 Tb

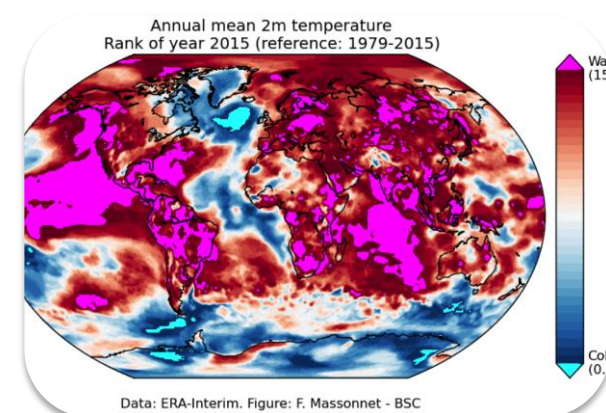
NMMB/BS C	Resolución horizontal (tamaño de celda)	Tamaño de las salidas para un año de simulación (meteorología, aerosoles, química en fase gas)
Resolución estándar	10km	2.3 Pb
Alta resolución	4km	9.1 Pb
Muy alta resolución	1km	36,5 Pb



“Data streaming”
para el pronóstico
de la calidad del aire



Analíticas y HPC en
el estudio del clima



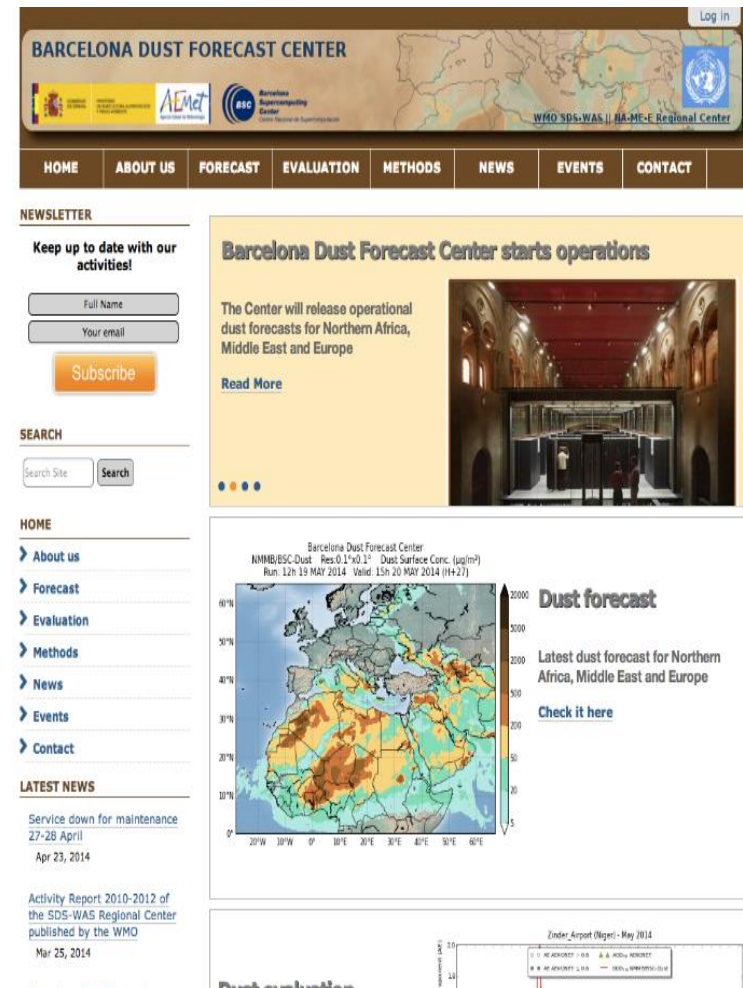
Analíticas como
servicio (AaaS)

Barcelona Dust Forecast Center (BDFC) y Sand and Dust Storm-Warning and Advisory System (SDS-WAS) for North Africa, Middle East and Europe, desarrollados conjuntamente entre BSC-CNS y AEMET:

- BDFC es el primer centro de la WMO especializado en predicción del polvo mineral y genera pronósticos para WMO GTS, EumetCAST y AEMET.
- SDS-WAS NAMEE es un proyecto de investigación y desarrollo multi-modelo agrupando salidas de modelos, observaciones de la red AERONET y datos de satélite.

<http://dust.aemet.es>

<http://sds-was.aemet.es>



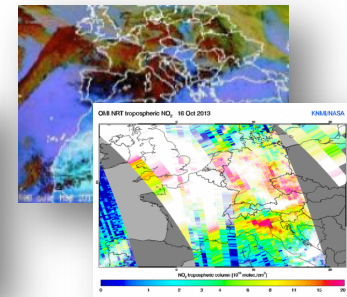
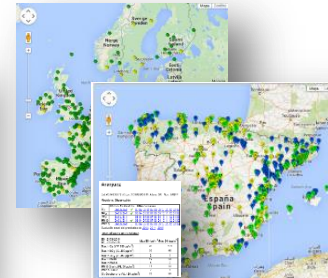
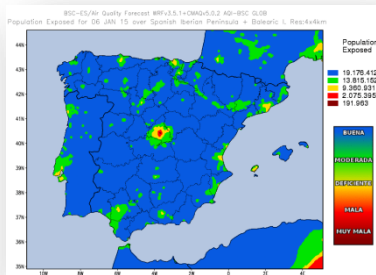
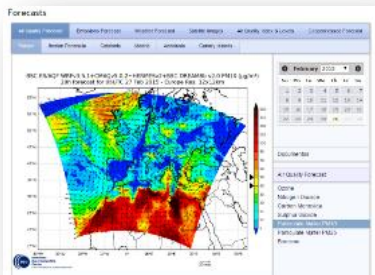
The screenshot shows the official website of the Barcelona Dust Forecast Center (BDFC). The header includes the BSC logo and the text 'Barcelona Supercomputing Center Centro Nacional de Supercomputación'. The main navigation bar contains links: HOME, ABOUT US, FORECAST, EVALUATION, METHODS, NEWS, EVENTS, and CONTACT. A 'Log in' link is also present. Below the navigation bar, there is a 'NEWSLETTER' section with a 'Subscribe' button. A 'SEARCH' bar is also available. The main content area features a large banner titled 'Barcelona Dust Forecast Center starts operations' with a photograph of the center's interior. Below this, there is a 'Dust forecast' section showing a map of North Africa, the Middle East, and Europe with dust concentration levels. The map is titled 'Dust forecast' and includes the text 'Latest dust forecast for Northern Africa, Middle East and Europe' and a link 'Check it here'. At the bottom, there is a 'LATEST NEWS' section with two entries: 'Service down for maintenance 27-28 April Apr 23, 2014' and 'Activity Report 2010-2012 of the SDS-WAS Regional Center published by the WMO Mar 25, 2014'.

AQF CALIOPE

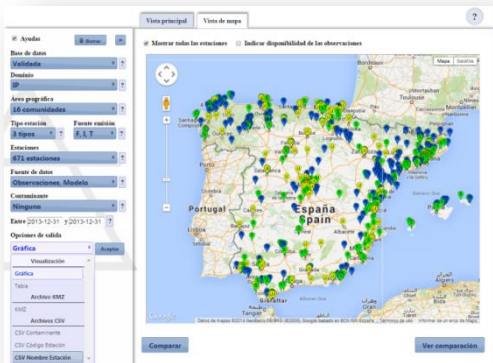
Sistema de pronóstico

Pronósticos

Pronóstico diario para la **meteorología, emisiones y calidad del aire**: Europa (12km), Península Ibérica (4km), Andalucía, Cataluña y Madrid (1km), desde 2007



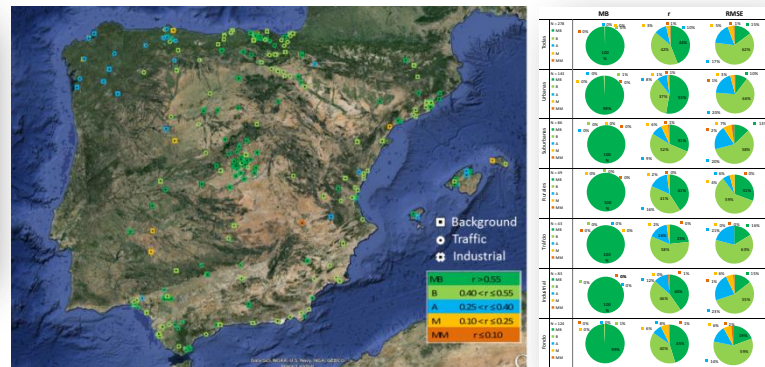
Bases de datos



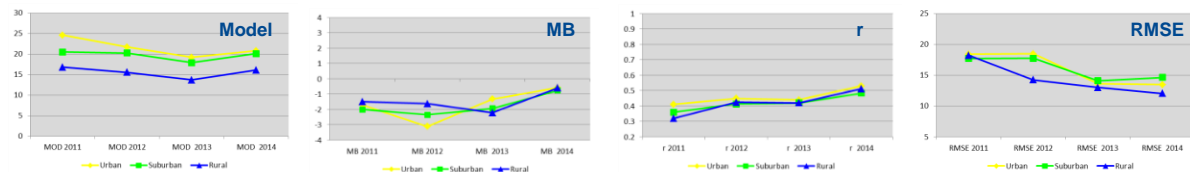
Evaluación en tiempo real



Evaluación anual con datos de estaciones

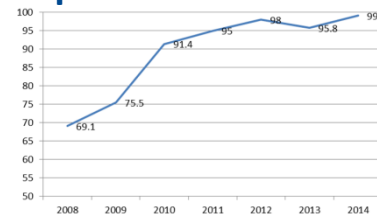


Seguimiento anual



PM10_KF evolución de la calidad del pronóstico (2011-2014)

Porcentaje de simulaciones completadas con éxito (%)





Apps

Categories ▾

Home

Top Charts

New Releases

My apps

Shop

Games

Family

Editors' Choice

CALIOPE

De Barcelona Supercomputing Center – Centro Nacional de Supercomputación

Abre iTunes para comprar y descargar Apps.

[Ver más de este desarrollador](#)



[Ver en iTunes](#)

+ Esta App se ha desarrollado tanto para iPhone como para iPad

Gratis

Categoría: [Salud y forma física](#)

Actualizado: 07/04/2014

Versión: 1.5.1

Tamaño: 16.2 MB

Idiomas: Alemán, Inglés

Desarrollador: Barcelona

Supercomputing Center –

Centro Nacional de

Supercomputación

© 2013 Barcelona

Supercomputing Center –

Centro Nacional de

Descripción

Esta aplicación permite visualizar el pronóstico de la calidad de aire en España para la hora siguiente a la de la consulta, además de los mapas de concentración para las próximas 12 horas, de los principales contaminantes atmosféricos sobre la península Ibérica.

[Soporte técnico de CALIOPE](#)

[...Más](#)

Lo nuevo en la versión 1.5.1

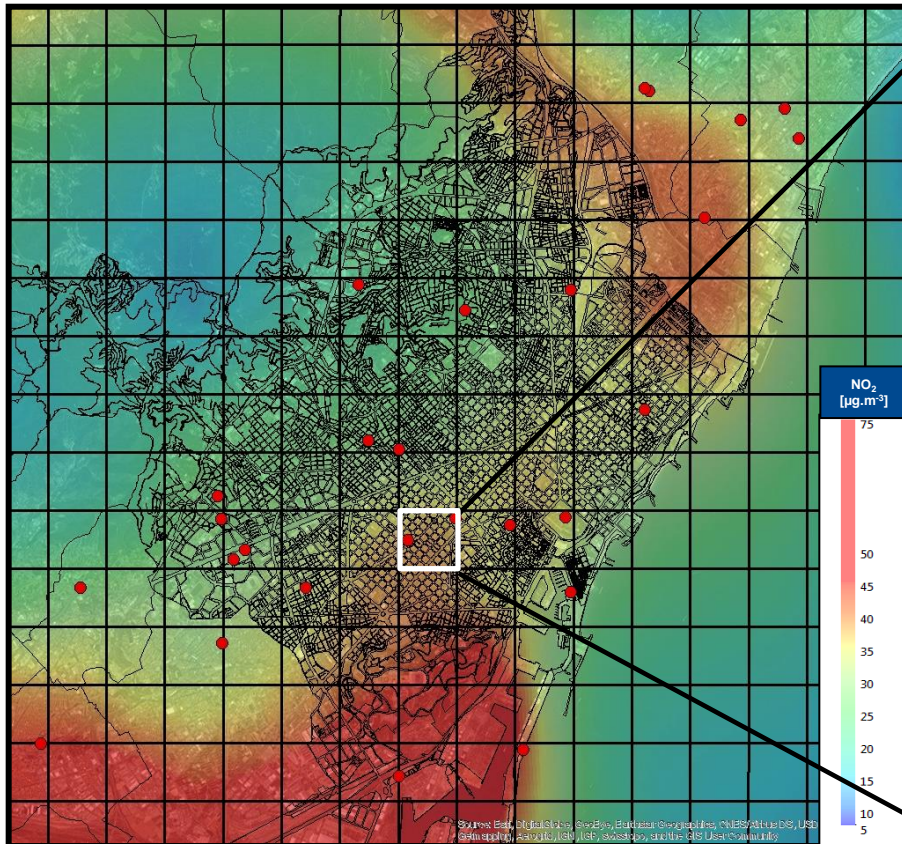
– Esta actualización soluciona algunos errores observados y realiza pequeños cambios visuales.

Capturas de pantalla

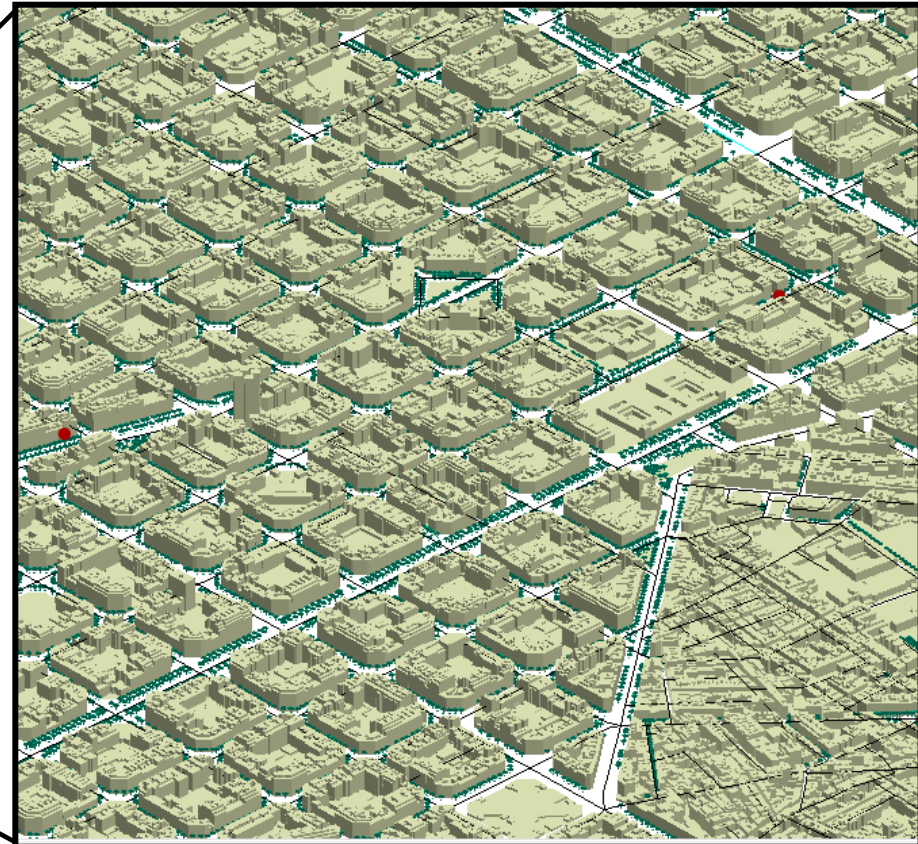
iPhone | iPad



Donde estamos ahora



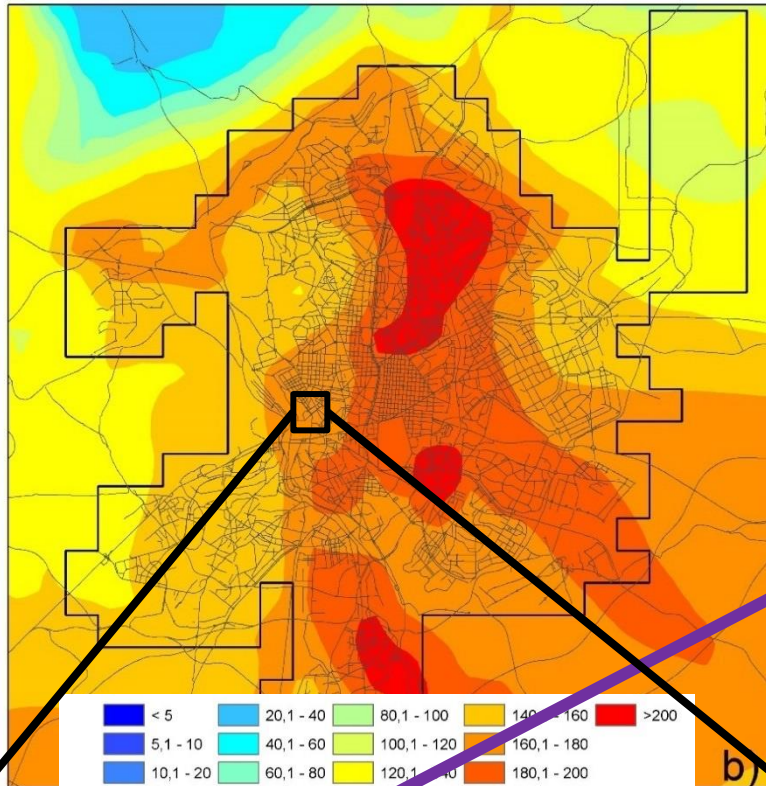
Donde queremos estar



Objetivo	Desarrollar un modelo de calidad del aire a escala urbana, con el uso de tecnologías Big Data para evaluar la calidad del aire urbano.
----------	--

Necesitamos más detalle

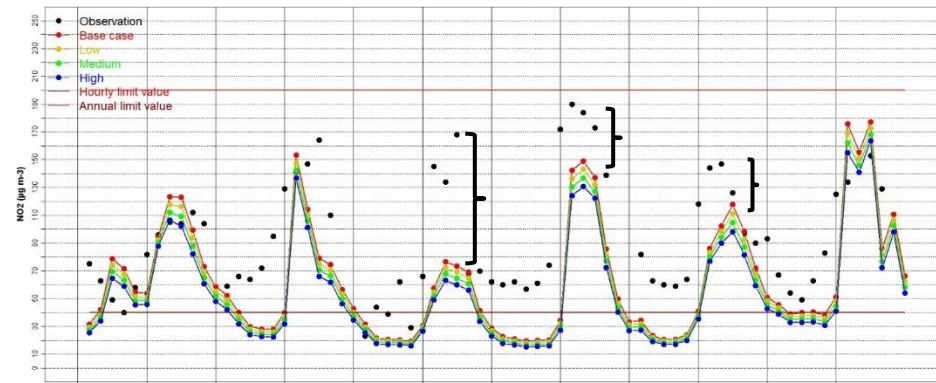
NO₂ (ug m⁻³) Max h
Base case; Madrid



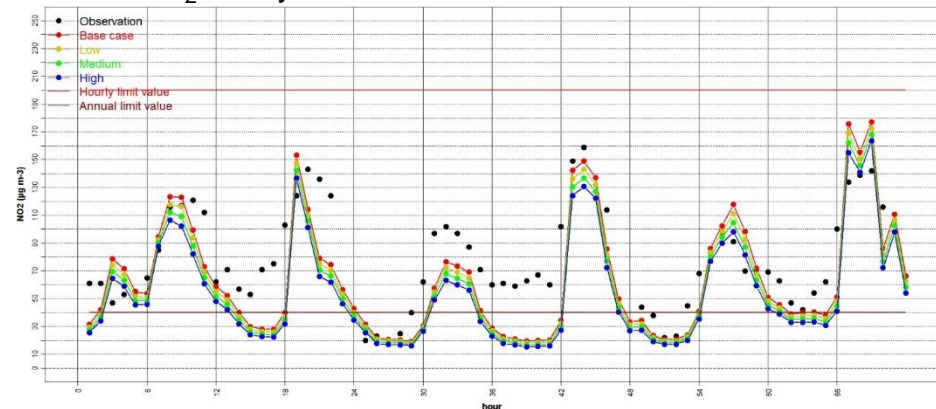
1) Los modelos mesoescalares (soluciones actuales) simulan correctamente las concentraciones de fondo (e.g. Plaza del Carmen).

2) No obstante no son capaces de reproducir las concentraciones a nivel de calles anchas (e.g. Plaza de España situada cerca de la Gran Vía)

NO₂ hourly concentration. Plaza de España station

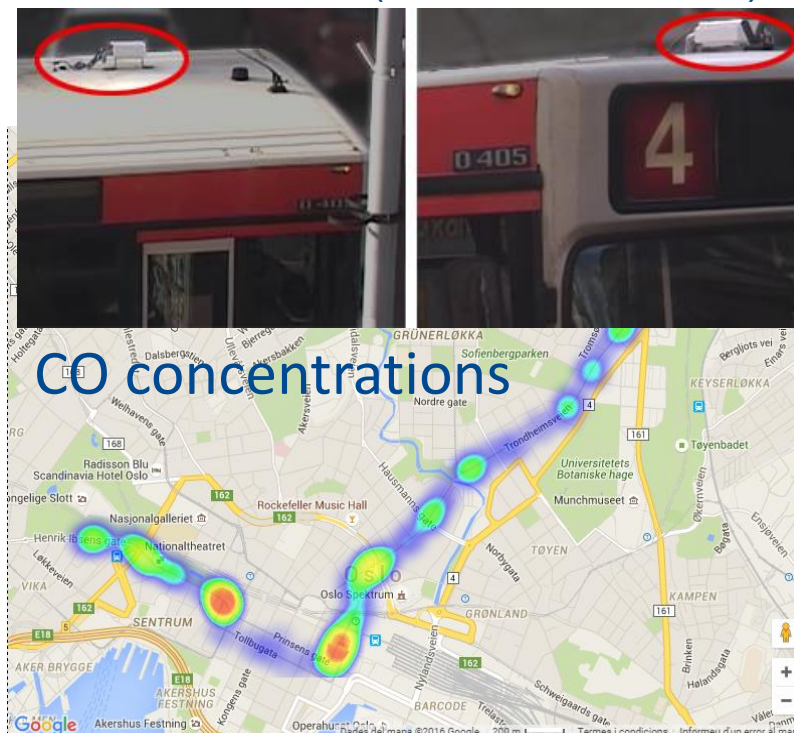
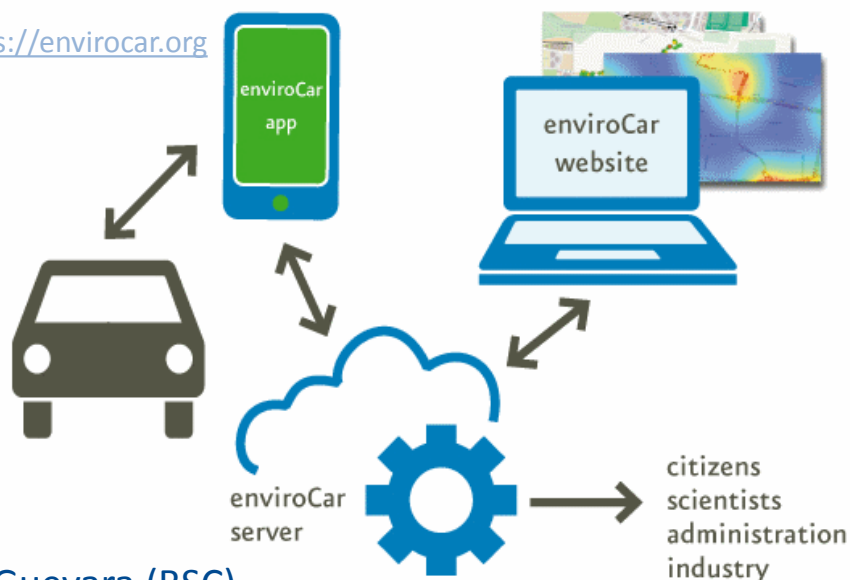


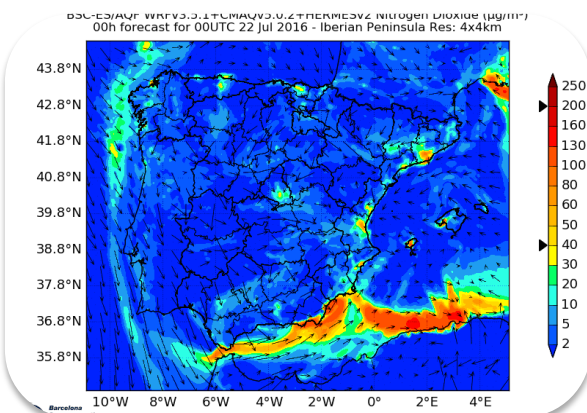
NO₂ hourly concentration. Plaza del Carmen station



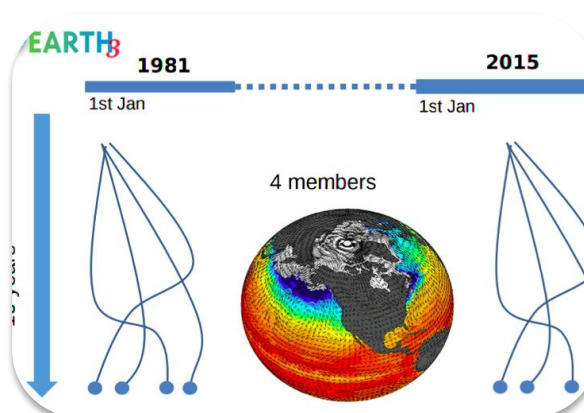
- La modelización de la calidad del aire en alta resolución requiere más datos de emisiones:
 - Recogida y tratamiento de los datos generados por sensores para alimentar un modelo de emisiones en tiempo real (y para validar las predicciones de calidad del aire)
 - Proporcionar los sensores con la tecnología adecuada es un gran reto
 - La gestión de grandes volúmenes es otro: el muestreo de 10 Hz, diez variables ~ 30MB / día, en toda la ciudad ~ 300 GB / día (10.000 vehículos)

<https://envirocar.org>

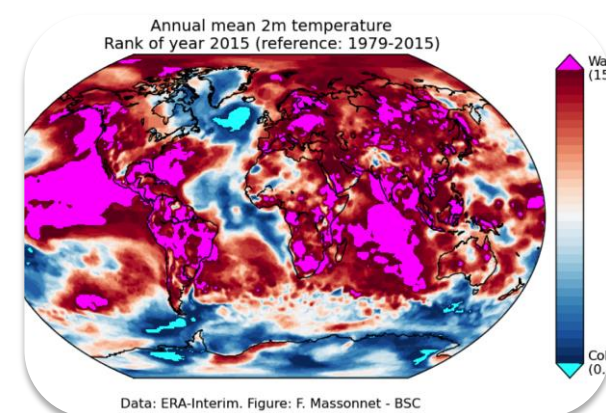




“Data streaming”
para el pronóstico
de la calidad del aire

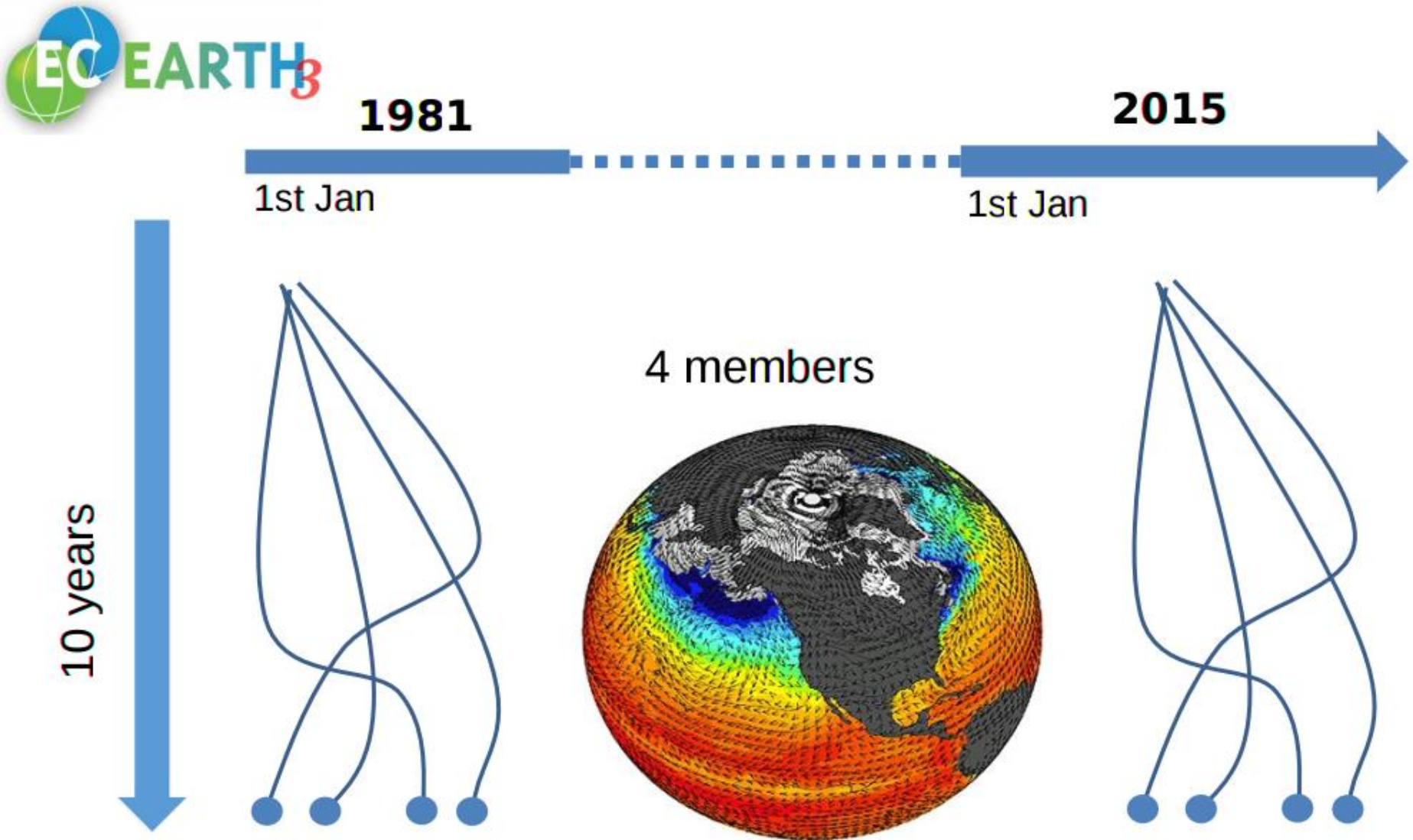


Analíticas y HPC en
el estudio del clima



Analíticas como
servicio (AaaS)

Experimento decadal



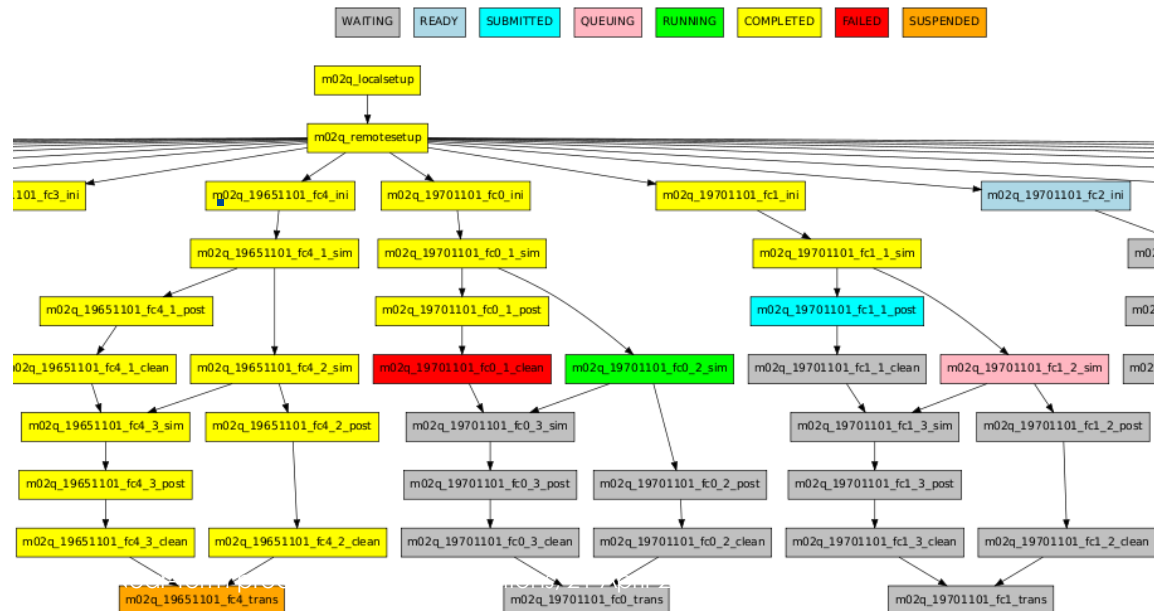
La predicción del clima permite ejecutar trabajos de forma independiente y simultánea lanzando miembros por conjuntos para diferentes fechas de inicio. Paralelizar este proceso no es trivial.

Se requiere un “workflow manager” potente y fiable.

EC-Earth3 at Lindgren, PDC						
Number of Start Dates		1	5	10	10	20
Number of Members		1	5	5	10	10
Number of Independent Simulations		1	25	50	100	200
T159-ORCA1	Cores	144	3600	7200	14400	28800
	Wall-clock Time (Hours) / year	5	5	5	5	5
	CPU Time (Hours) / year	720	18000	36000	72000	144000
	Output Size (GB) / year	10,80	480	960	1920	3840
T255-ORCA1	Cores	360	9000	18000	36000	72000
	Wall-clock Time (Hours) / year	5	5	5	5	5
	CPU Time (Hours) / year	1800	45000	90000	180000	360000
	Output Size (GB) / year	19,20	5184	10368	20736	41472
T799-ORCA025	Cores	1104	27600	55200	110400	220800
	Wall-clock Time (Hours) / year	40	40	40	40	40
	CPU Time (Hours) / year	44160	1104000	2208000	4416000	8832000
	Output Size (GB) / year	256,80	6420	12840	25680	51360

- **Automatización:** La preparación y ejecución, post-proceso y transferencia de las salidas, lo gestiona todo Autosubmit. El usuario no debe intervenir.
- **Identificación:** Asigna identificadores únicos para cada experimento y almacena la información sobre la versión del modelo, opciones de configuración, etc.
- **Tolerancia a fallos:** los reintentos automáticos y la capacidad de repetir trabajos en el caso de encontrar datos dañados o perdidos.
- **Versatilidad:** actualmente ejecuta EC-Earth, los modelos NEMO y NMMB en varias plataformas.

Workflow de un experimento monitorizado con Autosubmit
(amarillo = completed, verde = running, rojo = failed, ...)

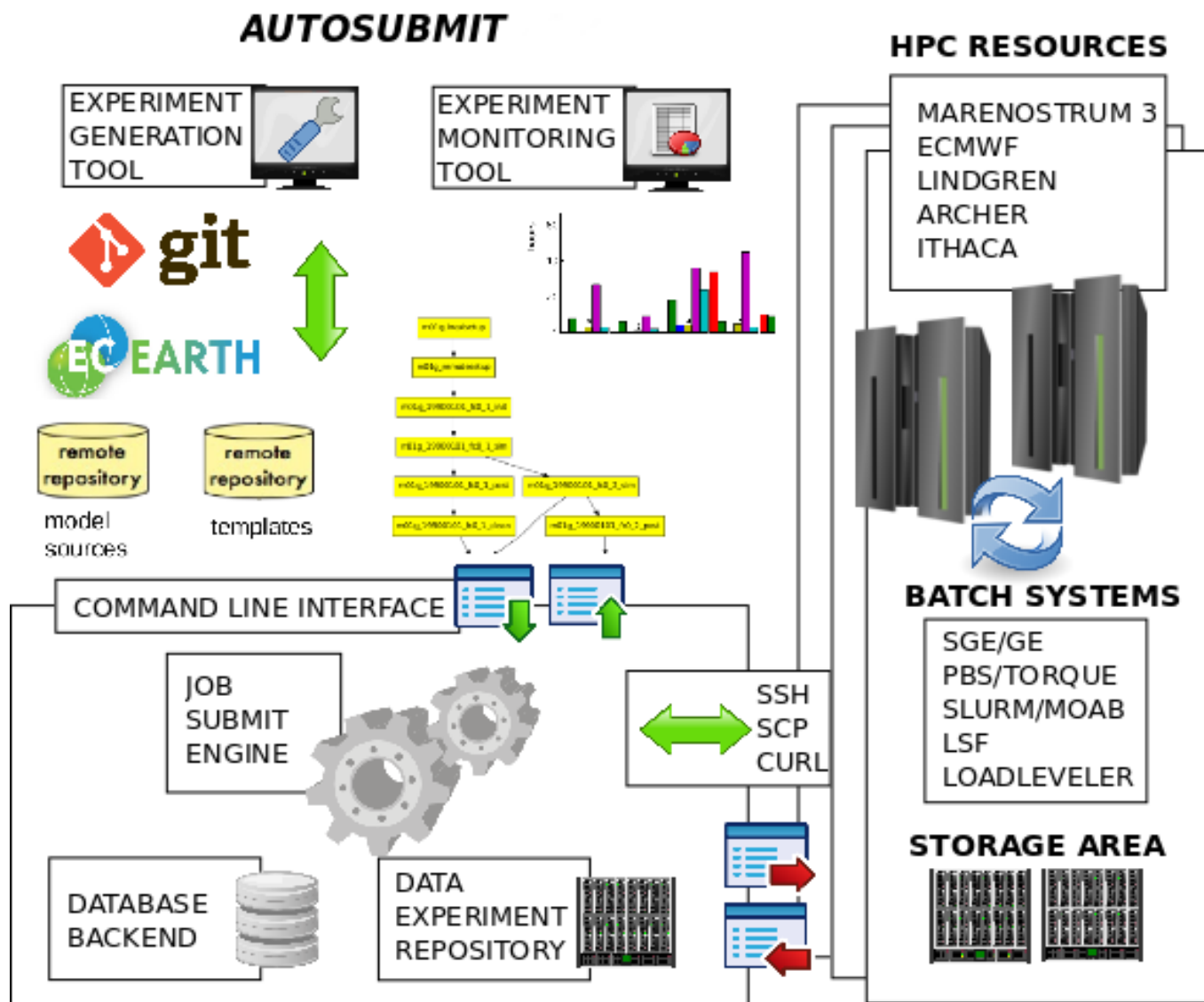


Workflow Manager: Autosubmit



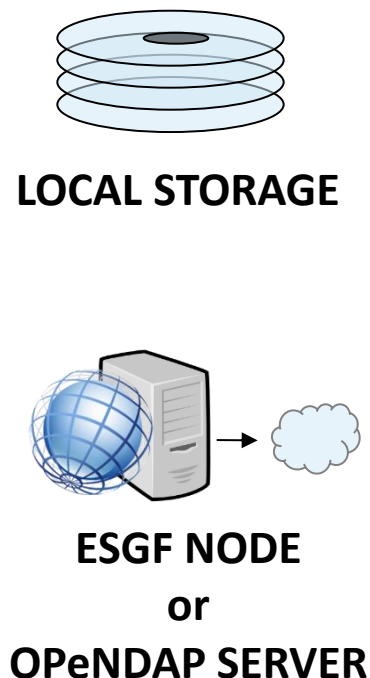
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

EXCELENCIA
SEVERO
OCHOA



<https://pypi.python.org/pypi/autosubmit/>

S2dverification es un paquete de R para verificar los pronósticos decadales mediante la comparación de las simulaciones con datos de observación. Permite el análisis de los datos disponibles de forma local o remota. También se puede utilizar online, mientras se ejecuta el modelo.



s2dverification package

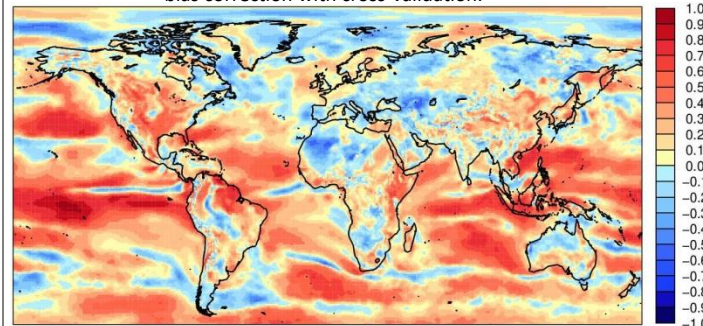
- Supports datasets stored locally or in ESGF (OPeNDAP) servers.
- Exploits multi-core capabilities
- API available
- Collects observational and experimental datasets stored in multiple conventions:
 - NetCDF3, NetCDF4
 - Supports specific folder and file naming conventions.

BASIC STATISTICS

SCORES EXTREMES

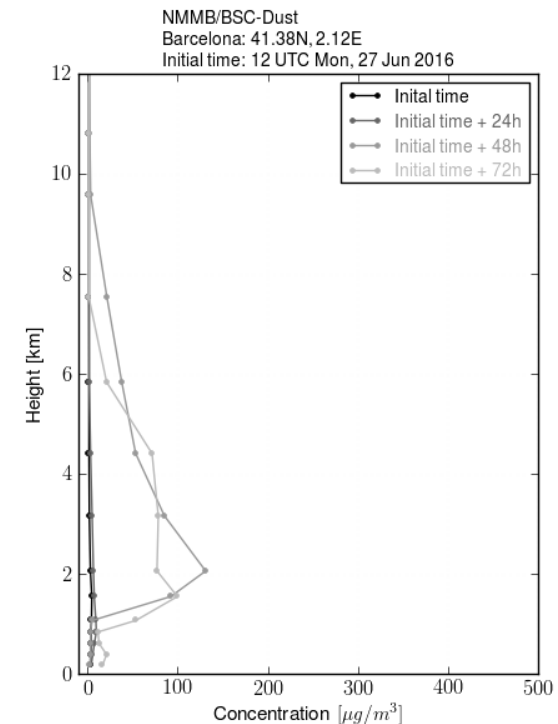
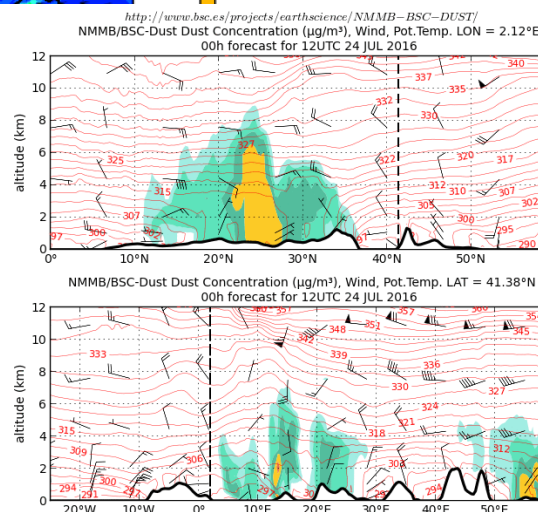
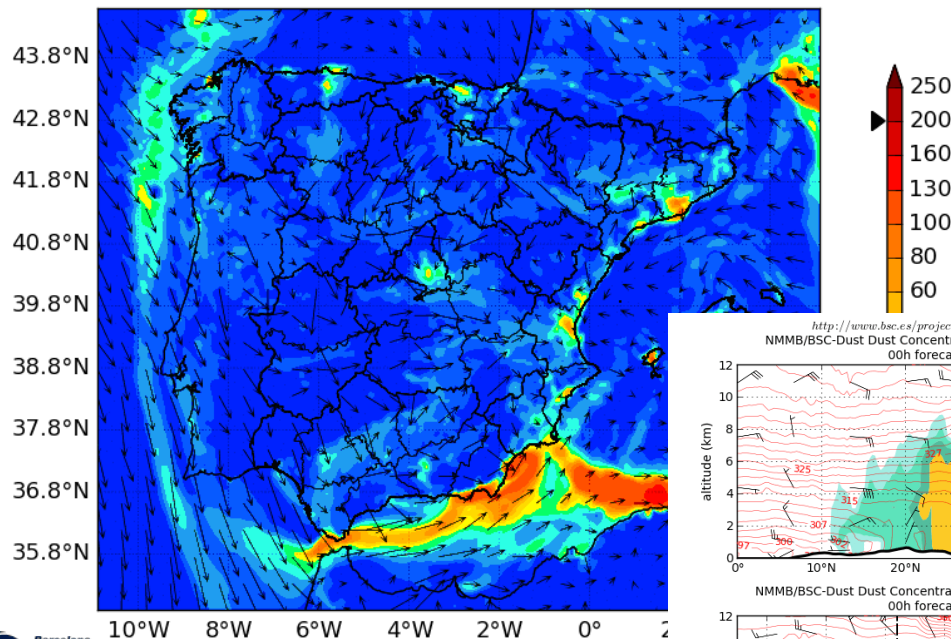
PLOTS

Anomaly Correlation Coefficient. 10M Wind Speed ECMWF S4 1 month lead with start dates once a year on first of November and Era-Interim in DJF from 1981 to 2011. Simple bias correction with cross-validation.



- Un módulo de Python, MapGenerator para la visualización de la calidad del aire
 - Python y GrADS

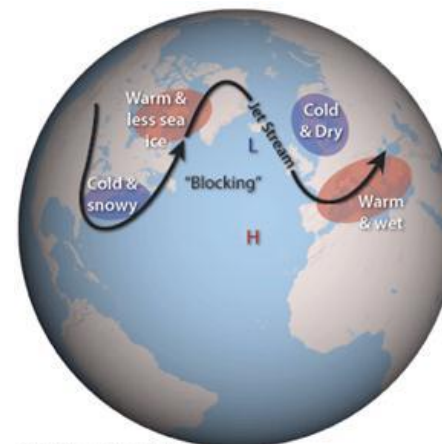
BSC-ES/AQF WRFv3.5.1+CMAQv5.0.2+HERMESv2 Nitrogen Dioxide ($\mu\text{g}/\text{m}^3$)
00h forecast for 00UTC 22 Jul 2016 - Iberian Peninsula Res: 4x4km



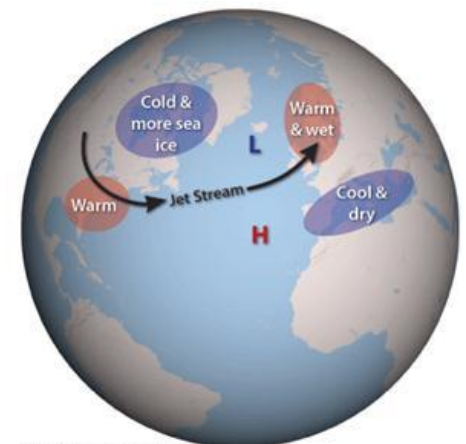
¿Qué es un diagnóstico?



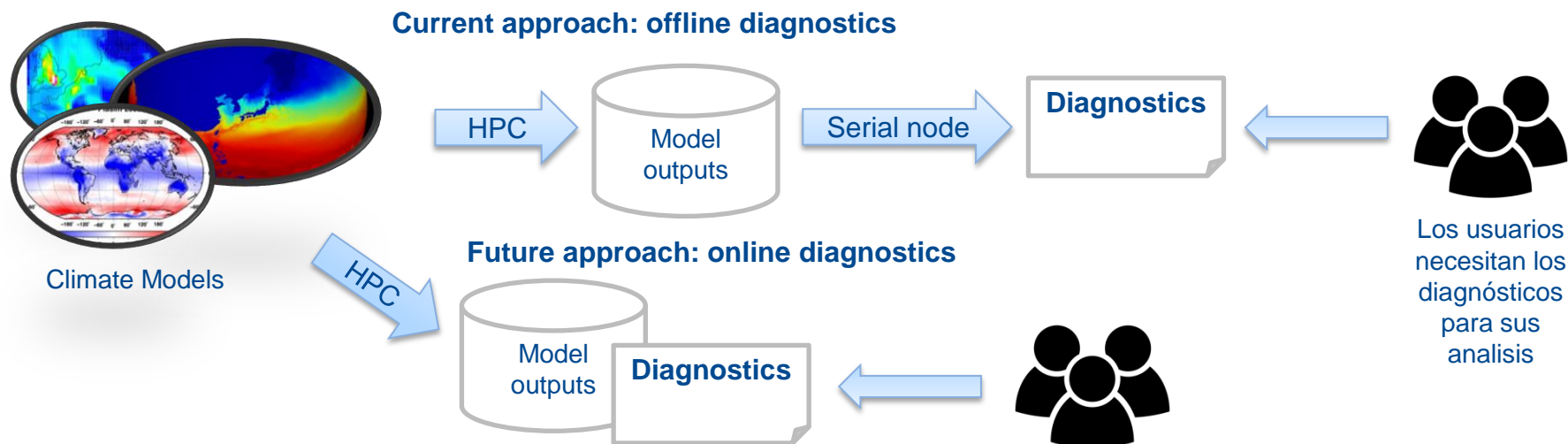
- PRIMAVERA producirá 600 TB de datos en dos años
- Los usuarios trabajan con una colección de índices que resumen la calidad de una simulación
- Estos índices se llaman **DIAGNÓSTICOS** o **METRICAS** y requieren data de alta resolución temporal, comparativas con observaciones.
- Los usuarios no pueden trabajar de una forma eficiente con todos los datos, de aquí que se necesiten estos valores



NAO Negative Mode

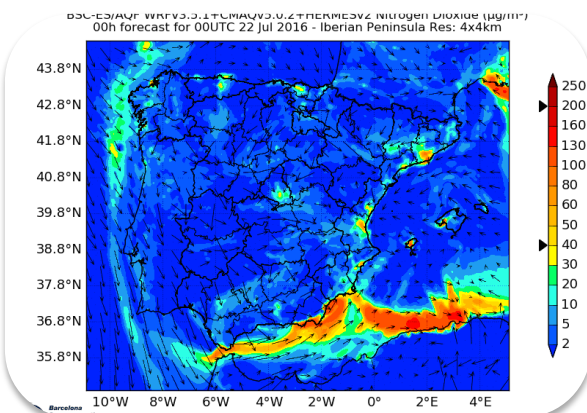


NAO Positive Mode

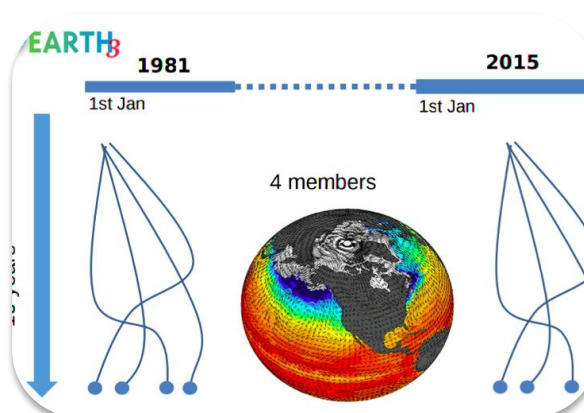


- Diagnósticos calculados como “Analytics as a Service”
 - Diagnosticos online (durante la ejecución del modelo)
 - Reducción del movimiento de datos
 - Diagnósticos calculados en los nodos de computación
 - Nuevos diagnósticos (data mining de extremos) posibles
 - El usuario recibe los datos antes → crucial para la adaptación al cambio climático y para desarrollar servicios climáticos (públicos y privados)

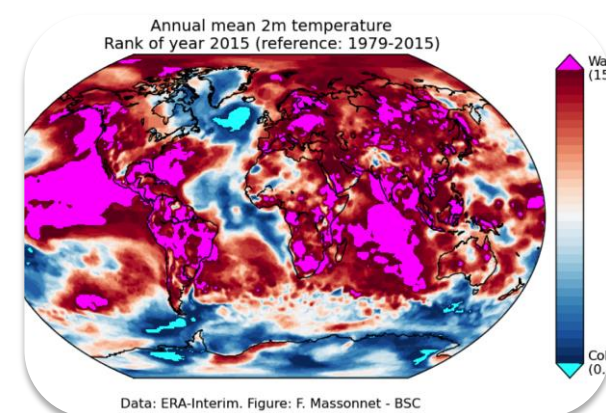
- Se requerirá una importante modificación de los códigos para correr de forma eficiente con las configuraciones en alta resolución (el trabajo en colaboración con Computer Science es indispensable)
- Los diagnósticos deberán ser implementados desde cero y adaptados a correr en paralelo en más de un nodo del servidor de escritura
- Es un esfuerzo coordinado con otras instituciones que utilizan aplicaciones similares o que tienen problemas similares:
 - RIKEN: investigación para reducir el tiempo para calcular los diagnósticos en una simulación a muy alta resolución (200 días para procesar una simulación global de 960m de resolución)
 - ARGONNE: trabajando con el mismo problema en el desarrollo del modelo ACME para el DoE



“Data streaming”
para el pronóstico
de la calidad del aire



Analíticas y HPC en
el estudio del clima



Analíticas como
servicio (AaaS)

El cambio climático está ocurriendo

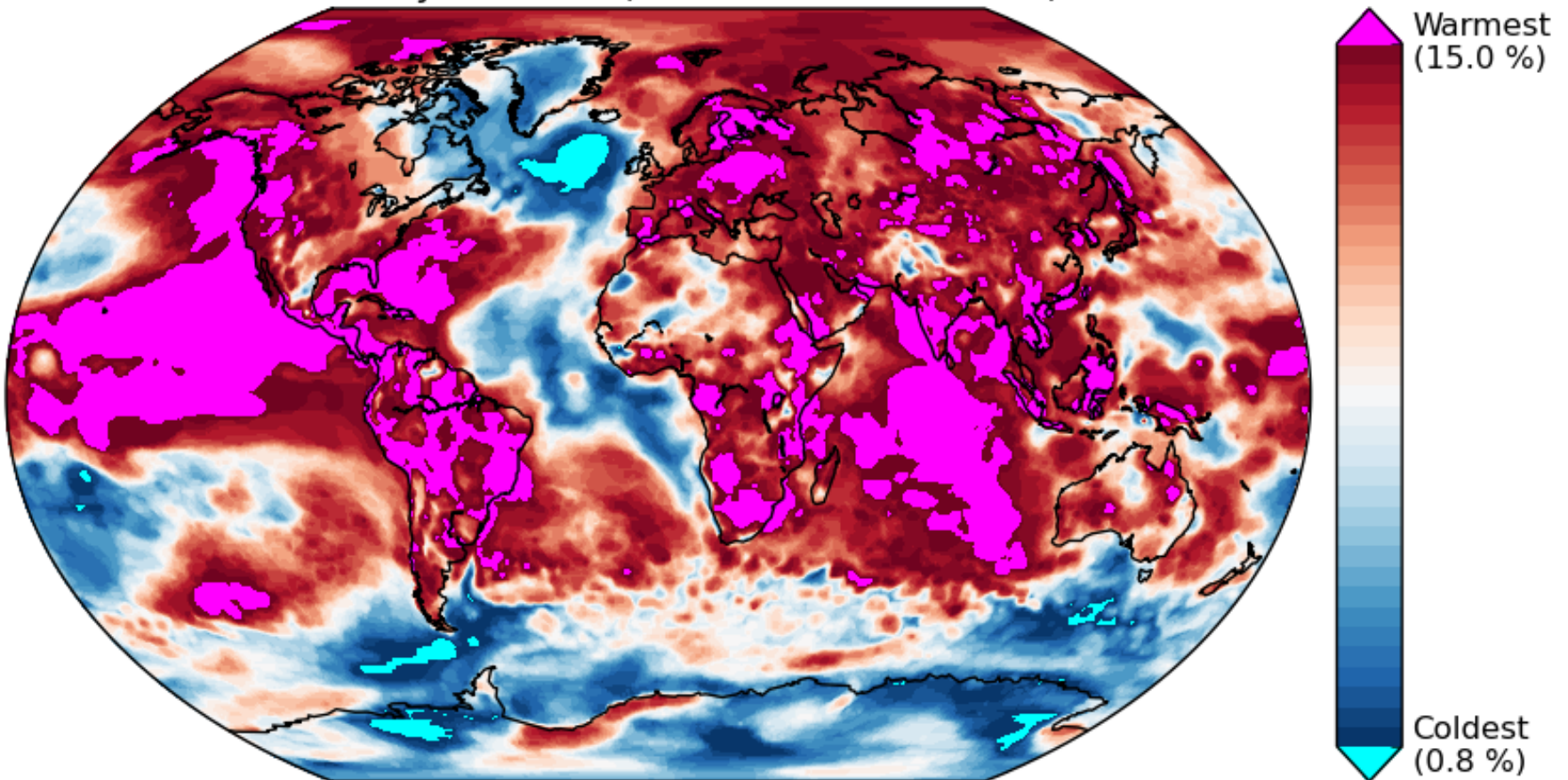


Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

EXCELENCIA
SEVERO
OCHOA

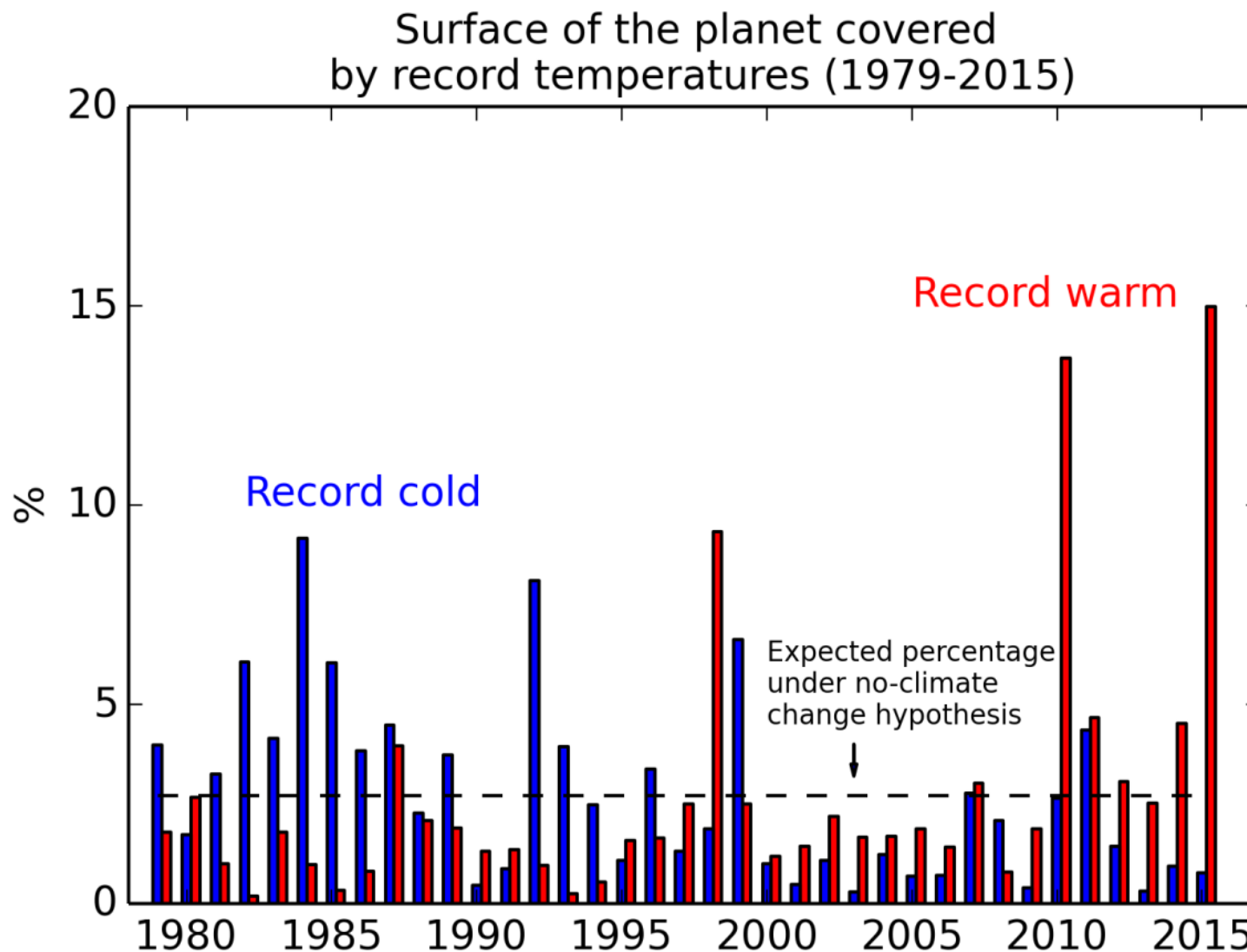
Ranking de la temperatura media anual de 2015 en los últimos 37 años utilizando los datos de ERA-Interim.

Annual mean 2m temperature
Rank of year 2015 (reference: 1979-2015)



Data: ERA-Interim. Figure: F. Massonnet - BSC

El cambio climático está ocurriendo

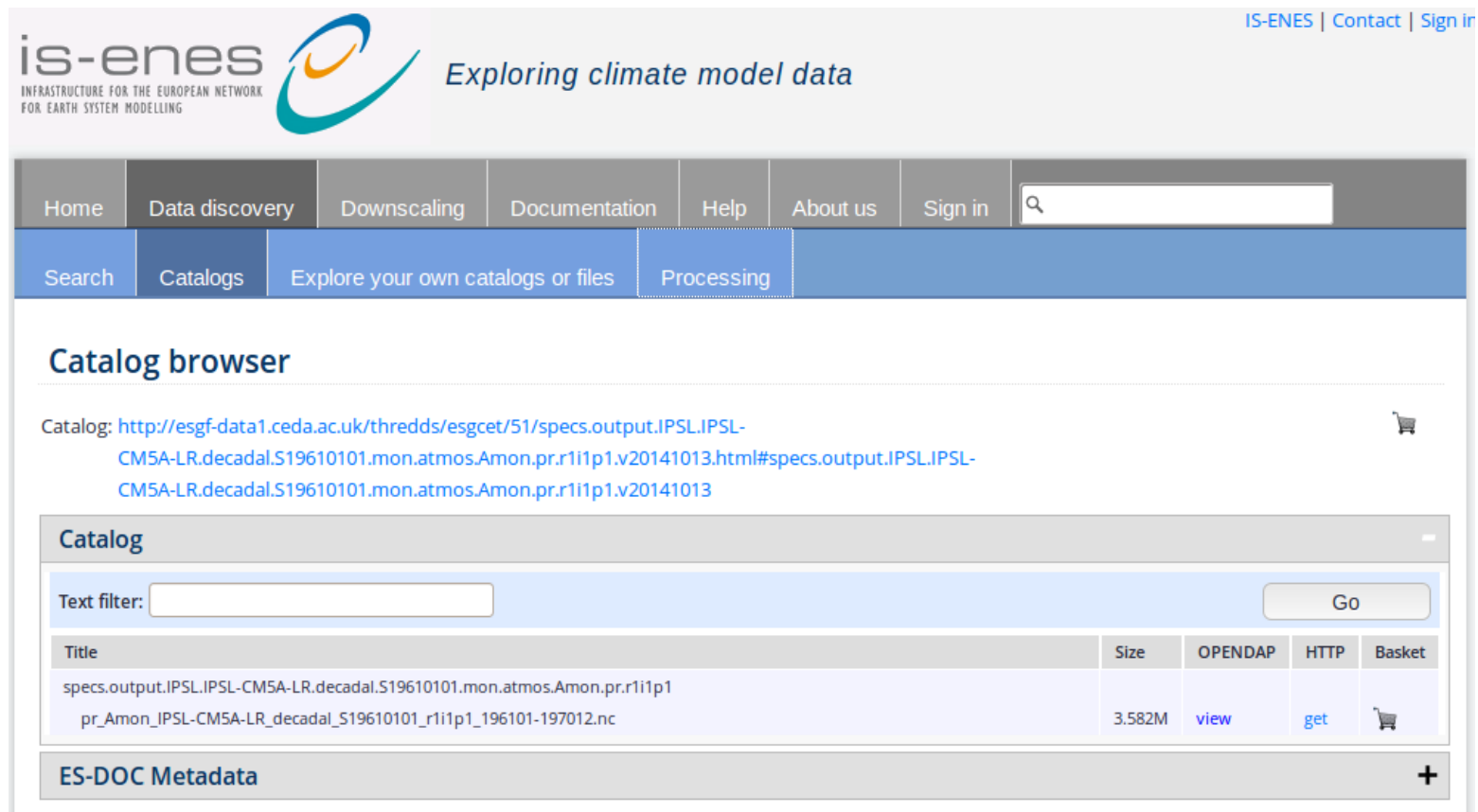


Los usuarios necesitan servicios que respondan a sus preguntas

Un ejemplo de este tipo de servicios es <http://climate4impact.eu>

Es un aspecto clave para el éxito de nuestra investigación

Basado en el acceso a **ESGF**.



The screenshot shows the 'is-enes' catalog browser interface. The header includes the 'is-enes' logo (Infrastructure for the European Network for Earth System Modelling) and the tagline 'Exploring climate model data'. Navigation links for 'IS-ENES', 'Contact', and 'Sign in' are in the top right. A main navigation bar contains 'Home', 'Data discovery', 'Downscaling', 'Documentation', 'Help', 'About us', and 'Sign in'. Below this is a secondary bar with 'Search', 'Catalogs', 'Explore your own catalogs or files', and 'Processing'. The 'Catalog browser' section displays a search result for a catalog URL. Below the search bar, a table lists the found data files with columns for Title, Size, OPENDAP, HTTP, and Basket.

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK
FOR EARTH SYSTEM MODELLING

Exploring climate model data


IS-ENES | Contact | Sign in

Home Data discovery Downscaling Documentation Help About us Sign in

Search Catalogs Explore your own catalogs or files Processing

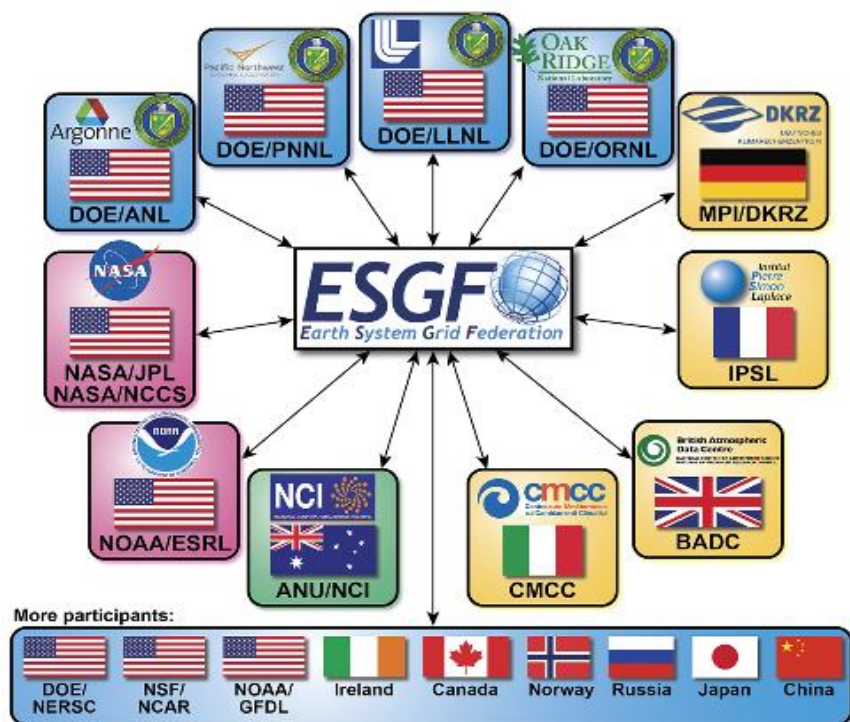
Catalog browser

Catalog: <http://esgf-data1.ceda.ac.uk/thredds/esgcat/51/specs.output.IPSL.IPSL-CM5A-LR.decadal.S19610101.mon.atmos.Amon.pr.r1i1p1.v20141013.html#specs.output.IPSL.IPSL-CM5A-LR.decadal.S19610101.mon.atmos.Amon.pr.r1i1p1.v20141013>

Title	Size	OPENDAP	HTTP	Basket
specs.output.IPSL.IPSL-CM5A-LR.decadal.S19610101.mon.atmos.Amon.pr.r1i1p1 pr_Amon_IPSL-CM5A-LR_decadal_S19610101_r1i1p1_196101-197012.nc	3.582M	view	get	

ES-DOC Metadata

- La Earth System Grid Federation es un esfuerzo de código abierto que proporciona una robusta plataforma de datos distribuida, lo que permite un acceso mundial a peta/exa datos meteorológicos y climáticos.
- ESGF promueve un formato común y proporciona herramientas de búsqueda e indexación de datos.

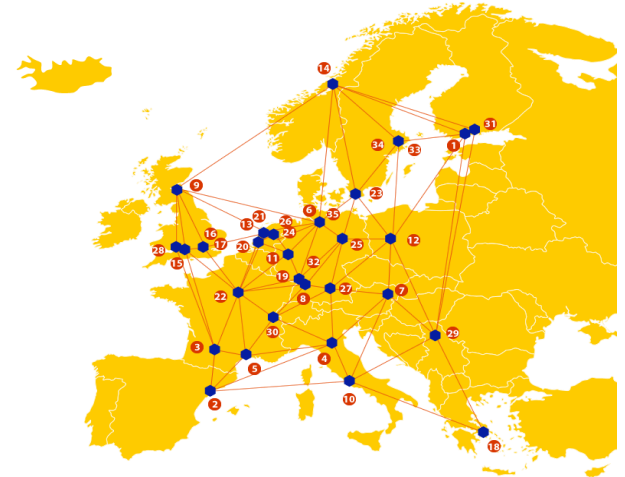


**Additional participants could not be illustrated in this figure.*

- ESGF está organizado como una red mundial de nodos distribuidos
 - Nodos de datos (THREDDS server, gridFTP server, ...)
 - Nodos de índices (descripción de la base de datos)
 - Nodos de identidad (autenticación de usuarios)
 - Nodos de computación (análisis y visualización de datos)

- Piloto EUDAT

- “Support to scientific research on seasonal-to-decadal climate and air quality modelling”
 - *Data sync & exchange*
 - *Data replication*
 - *Data discovery & search*
 - *Data repository & sharing*
 - *Data staging*



- RDA grupo de interés en “Weather, climate and air quality”
 - Abordar los retos para el uso y el análisis de los grandes volúmenes de datos de la comunidad
 - Tratar problemas comunes con otras comunidades científicas
 - Hacer visible este tipo de problemáticas
 - Mailing list: earthsciences-rda-ig@bsc.es



¿Cómo empezar?



**Barcelona
Supercomputing
Center**
Centro Nacional de Supercomputación

EXCELENCIA
SEVERO
OCHOA

- <https://www.kaggle.com/>

The screenshot shows the Kaggle website interface. At the top, there's a navigation bar with links for Competitions, Datasets, Kernels, Forums, and Jobs. A 'Sign Up' button and a 'Log In' link are also present. The main header features the Kaggle logo and the competition title 'Climate Change: Earth Surface Temperature Data' with a subtitle 'Exploring global temperatures since 1750'. Below the title, it says 'by Berkeley Earth · last updated 4 months ago'. There are tabs for Overview, Kernels, Discussion, Activity, Download (84 MB), New Notebook, and New Script. The 'Overview' tab is selected. Below the tabs, there are three columns: Kernels, Discussion, and Top Contributors. The Kernels column lists three notebooks: 'New York, New York...' (19 votes), 'Continental US Climate C...' (12 votes), and 'Are Temperatures Normal?' (11 votes). The Discussion column lists three discussions: 'My first kaggle notebook' (0 replies), 'Indian cities temperture ho...' (5 replies), and 'Climate Change Analysis' (0 replies). The Top Contributors column lists three users: Ben Watson (1st), SvenBrüssow (2nd), and Laura Jonutyte (3rd). At the bottom, there's a 'Recent Activity' section.

Climate Change: Earth Surface Temperature Data
Exploring global temperatures since 1750

by **Berkeley Earth** · last updated 4 months ago

Overview Kernels Discussion Activity Download (84 MB) New Notebook **New Script**

Kernels

New York, New York... run 4 months ago	19 votes
Continental US Climate C... run 4 months ago	12 votes
Are Temperatures Normal? run 4 months ago	11 votes

Discussion

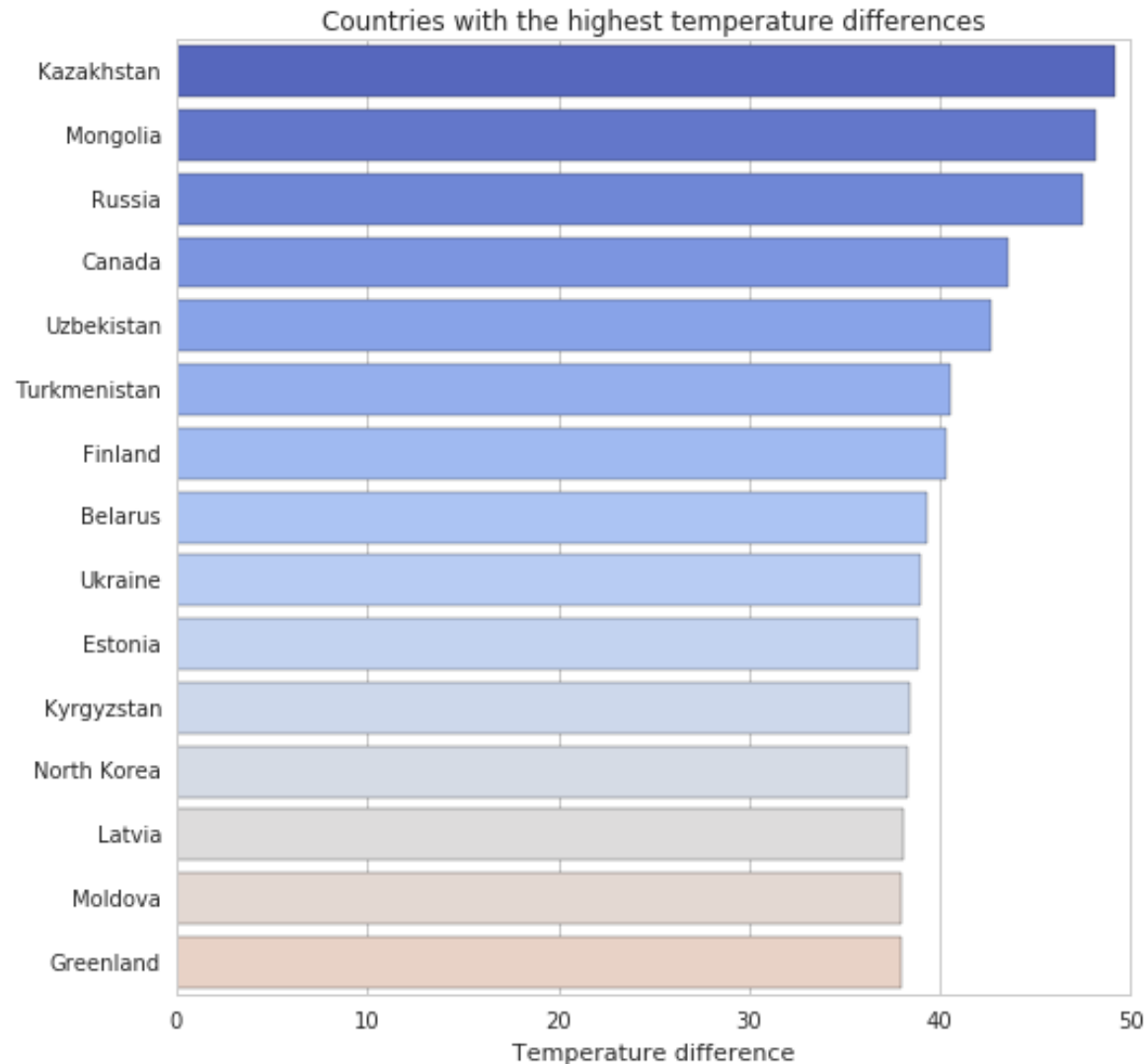
My first kaggle notebook 1 day ago	0 replies
Indian cities temperture ho... 3 days ago	5 replies
Climate Change Analysis 1 week ago	0 replies

Top Contributors

	Ben Watson	1st
	SvenBrüssow	2nd
	Laura Jonutyte	3rd

Recent Activity

<https://www.kaggle.com/dmitriy19/d/berkeleyearth/climate-change-earth-surface-temperature-data/some-visualizations-for-climate-change>



Preguntas



**Barcelona
Supercomputing
Center**
Centro Nacional de Supercomputación



0 1 0 1 0
1 0 1 0 1
0 1 0 1 0
1 0 1 0 1
0 1 0 1 0
1 0 1 0 1
0 1 0 1 0
1 0 1 0 1
0 1 0 1 0
1 0 1 0 1
0 1 0 1 0
1 0 1 0 1
0 1 0 1 0



**Barcelona
Supercomputing
Center**

Centro Nacional de Supercomputación



¡Gracias!

kim.serradell@bsc.es